Many High-Quality Randomized Controlled Trials in Sports Physical Therapy Are Making False-Positive Claims of Treatment Effect

Many high quality RCTs in sports physical therapy are making false positive claims of treatment effect: a systematic survey

[1,2] Bleakley Chris (0000-0001-9032-9691); [2] Reijgers J; [2] Smoliga James, M

1. School of Health Science, Faculty of Life and Health Science, Ulster University, Newtownabbey, Northern Ireland, BT370QB

2. Department of Physical Therapy, Congdon School of Health Science, High Point University, 1 University Parkway, High Point, NC 27260

Chris Bleakley associate professor Jonathan Reijgers spt James Smoliga professor

Correspondence address as above

Corresponding author: Chris Bleakley

Email c.bleakley@ulster.ac.uk

Word count: 2948

References: 32

Figures: 3

Key words: Null hypothesis significance testing, Randomised controlled trials, False Positive Risk, p-values

Abstract

Objective: To examine the risk of false positive reporting within high quality randomized controlled trials (RCTs) in the sports physical therapy field.

Design: Cross-sectional

Methods: We searched the PEDro database for parallel design 2-arm RCTs reporting positive treatment effects based on null hypothesis significance testing, and scoring >6/10 on the PEDro scale. No restrictions were made on pathology, intervention or outcome variables. Sixty-two of 212 RCTs reported positive effects in at least one outcome variable. We estimated False Positive Risk (FPR) using the FPR Web Calculator (version 1.5) based data on: *n* of participants, p-value, and effect size. For each study, FPR was estimated using a range of prior probability assumptions: 0.2 (skeptical hypothesis), 0.5 and 0.8 (optimistic hypothesis).

Results: We calculated the FPR associated with 189 statistically significant findings ($p<0.05$) reported across 44 trials. The median FPR was 9% (25^{th}-75^{th} PCTL: 2-22%). 59% of statistically significant results (102/174) had FPR >5%, and 16% (28/174) had FPR >50%. Changing the prior probability from skeptical to optimistic reduced the median FPR from 30% (25^{th}-75^{th} PCTL: 9-54%) to 2% (25^{th}-75^{th} PCTL: 0.5-7%).

Conclusion: High quality RCTs using null hypothesis significance testing often overestimated treatment effects. The median false positive risk (FPR) was 9% -- in one in 10 trials, the researchers falsely concluded there was a treatment effect. Future RCTs in sports physical therapy should be informed by pre study odds and a minimum FPR estimation.

Introduction

High quality research can help clinicians and patients decide which treatments are likely to be most effective.[15] Successful replication of research findings is an integral part of the scientific process, and represents a more robust evidence base for clinical decision making. However, there is concern that the majority of published research claims are false.[17]

In a survey of 1576 researchers, more than 70% had tried and failed to reproduce another scientist's experiment, and more than half failed to reproduce their own experiments.[1] In preclinical research, only 11 - 49% of research findings have been successfully replicated, [10] with similar figures reported in psychological science.[27] Although evidence-based practice should substantially improve the quality and cost of healthcare, serious concerns regarding randomized controlled trial design and statistical analysis raise questions about the validity of evidence-based interventions.

Experimental analysis in medicine is usually frequentist: conclusions informed by p (probability) values generated from null hypothesis significance testing. However, many researchers and clinicians are unable to define or accurately interpret p-values.[5] Common misconceptions are that a p-value represents 'the probability that the results occurred by chance' or 'the probability that the null hypothesis (H0) is true'[5] or 'the probability that the hypothesis being tested is true.'[24] A p-value only represents the probability that the obtained data, or more extreme values, could be obtained if H0 is true[24] – the probability

of the data, on the condition that the null hypothesis is true. For more help understanding P values, see[18]

Misinterpreting the results of statistical tests makes it difficult to disentangle true from false positive findings. Understanding and accurately applying appropriate statistics defends against false discoveries.[24] Central, is quantifying the false positive risk (FPR) – "the probability of observing a statistically significant p-value and declaring that an effect is real, when it is not."[6] The FPR within different areas of biomedical science has been conservatively estimated at 25%.[24] This means that in at least 1 in 4 studies, the researchers falsely concluded a treatment effect. Others[4, 5, 17] have used data simulations to demonstrate experimental studies can carry a high FPR, even if their effect sizes are large and/or p-values are less than commonly used thresholds such as $p < 0.01$.

The issue of irreproducible data has been discussed by scientists for decades.[2] However it has received little attention in health care. No one has examined FPR using primary data extracted from high-quality clinical experimental research. Given the criticism of a weak evidence base for orthopedics and sports medicine,[3, 14, 22, 26] our objective was to estimate the false positive risk (FPR) of high-quality randomized controlled trials (RCTs) in sports physical therapy. Our secondary objectives were to examine the relationship between FPR and reported p-values by quantifying the number of studies with FPR >5%; and to determine how FPR changed based on assumptions around the prior probability of effect.

Methods

Trial selection

Trials were sourced from the Physiotherapy Evidence Base (PEDro), which is a freely accessible database aiming to "guide users to trials that are more likely to be valid" and "guide clinical practice."[19] In addition to serving as a database for clinical trials, PEDro includes a 10-item scale quantifying study quality.[14, 7]

We identified all RCTs scoring >6/10 and categorized in the subcategory of 'sports' (sports is defined by PEDro as "papers which specifically mention sports injuries as well as conditions which commonly affect sports people (eg, ligament repairs)." Eligible RCTs must have employed null hypothesis significance testing to determine evidence of effect and a parallel group design. No restrictions were made on pathology, intervention type or date of publication. We excluded RCTs with: healthy participants only; >2 intervention groups; cross over, cluster or pilot study designs.

Data extraction and management

We extracted the following data from all eligible trials: population, number of participants, primary diagnosis, intervention, comparison, outcome(s), allocation ratio, follow up time, p-value, effect size, trial registration number, and a priori power calculation.

We subgrouped the trials as either 1). Positive: the attainment of a dichotomous threshold of statistical significance ($p < 0.05$) in at least 1 outcome; or 2). Null: reporting no evidence of effect ($p > 0.05$).

For all trials that reported evidence of effect (Positive studies), we extracted additional data. First, we extracted details of between-group comparisons, making no restriction on outcome construct or follow-up time. If there was a between-group comparison with a positive statistically significant finding, we extracted the p-value, the number of participants in each group, and when possible, we calculated the corresponding effect size (Hedges g). If a trial reported a threshold of $p<0.05$, rather than an exact p-value, we assumed that the p-value was one decimal place below the threshold value (e.g. 0.049).

Estimating the false positive risk

We calculated FPR using the False Positive Risk Web Calculator (version 1.5)[23] For further details of the analysis script and simulated examples of FPR calculations see [5, 6]. Calculating FPR requires imputation of the prior probability that there is a real effect [P(H1)] for a given treatment. In all trials, we initially assumed that P(H1), was 0.5 – that there was a 50% probability a treatment intervention had a positive underlying effect before the trial was conducted.[4, 5]

We ran additional simulations based on extreme prior probabilities of P(H1) =0.2, where the chances of a positive effect are very small (a skeptical hypothesis), and P(H1)=0.8 where chances of effect are almost certain (an optimistic hypothesis). We also applied a reverse Bayesian approach:[5, 25] using observed p-values to determine the prior probability that would be required to achieve a FPR of 5%. In all cases FPR estimations were calculated using the p-equals method,[23] which is the probability of observing a statistically

significant finding that is due to chance for a single result, rather than trying to estimate the long term error rate (lifetime FPR).

We calculated FPR for primary and secondary outcomes where applicable. When trials included multiple outcome measures but did not clearly specify a primary outcome, we assigned a primary outcome based on the nature of the research question and the following definition:[28] 'a specific key measurement(s) or observation(s) used to measure the effect of experimental variables in a study. We examined the relationship between all reported p-values and the corresponding FPR using descriptive statistics, scatter and violin plots.

Results

There were 212 RCTs scoring >6/10 within the 'sport' subcategory on PEDro. Ninety trials were excluded for the following reasons: not parallel design (2 group) randomized controlled trial (n=56); healthy participants/no clinical outcomes (n=23); non-English language (n=9); abstract/full text not available (n=2).

We included 122 RCTs; 49% (n=60/122) reported a null finding, and 51% (n=62/122) reported positive effects from at least one outcome (Figure 1). Full trial details can be found in the Supplemental data file. There were few differences between the subgroups (positive vs null) in primary diagnoses and treatment interventions (Figure 1). The majority of RCTs included participants with tendinopathy (n=47 studies), musculoskeletal pain

(n=19 studies) or ligament/joint problems (n=21 studies). Electro-physical agents (n=48), rehabilitation (N=37) and manual therapy (n=17) were the most common interventions.

Insert Figure 1

Diagnosis and Primary Treatment*

False Positive Risk

In trials reporting positive effects (n=62), 67% compared two different physiotherapeutic approaches, and 33% used either sham or placebo controls. The mean sample size was n=57.3 (SD=35.2; range 16-172). Twenty-nine percent of trials (18/62) were prospectively registered; 64% (40/62) reported using *a priori* sample size calculation. The majority of sample size estimations included alpha (Type 1 error) and beta (Type 2 error) levels of 5% and 20% respectively; and the anticipated *a priori* effect size used was 0.9 on average (SD 0.4, range 0.2- 2.2).

We could not calculate FPR in 18 trials due to missing data. In the remaining 44 trials, we calculated FPR associated with 189 between-group comparisons reported as statistically significant. Lower p-values were associated with lower FPR (Figure 2). The mean FPR (based on prior probability of 0.5) was 25.2% (SD 34.3). As the data were not normally distributed, the median FPR of 9% is more representative of the data's central tendency (25^{th}-75^{th} percentile: 2-24%). Sixty-three percent of reported p-values (119/189) were associated with FPRs greater than 5%; 18% (35/189) had a FPR greater than 50%.

Using a reverse Bayesian approach, 57% (68/119) of statistically significant findings (primary or secondary outcomes) would require prior probabilities greater than 0.8, if FPRs of 5% were to be achieved. FPR patterns were similar when examining only primary outcomes, with mean and median FPRs of 22.9% (SD 36.1) and 5% (25^{th}-75^{th} percentile: 1-22%) respectively.

Insert Figure 2

P-value vs False Positive Risk

[Data relate to 189 positive effects reported from high quality RCTs (n=44); FPR based on a prior probability of 0.5; Dashed line = reference if p-value was equal to FPR.]

The lowest FPR occurred when the prior probability of effect was assumed as 0.8, with median risk of 2% (25^{th}-75^{th} percentile: 0.6-7%) (Figure 3). False positive risk increased when prior probabilities of 0.2 were assumed: median risk of 29% (25^{th}-75^{th} percentile: 9-56%).

Insert Figure 3

FPR based on 3 different prior probability levels [P(H1)=0.2, P(H1)=0.5; P(H1)=0.8]

[In all calculations, data relate to 189 positive effects reported from high quality RCTs (n=44)]

Discussion

We found that 63% of statistically significant findings ($p<0.05$) in the sports physical therapy literature generated FPRs greater than 5%. Repeated simulations of t-tests suggest that if one uses $p=0.05$ to conclude a discovery, one will be wrong at least 30% of the time.[4] False discoveries (claiming a treatment effect is real when it isn't) may be minimized through better understanding of the FPR. This is the first time that the healthcare literature has been audited to determine the FPR using primary data extracted from higher quality clinical experimental research. The median FPR was 9% (25^{th}-75^{th} percentile: 2-24%), suggesting that approximately one in every 10 trials in the sports physical therapy field have falsely concluded a treatment effect.

There have been a range of proposals to help minimize unsubstantiated claims of effectiveness in research. One option has been to lower p-values thresholds to $p\leq0.001$, to keep false discovery rates below 5%.[4] Recently the American Society of Statisticians released a number of recommendations aimed at improving use of null hypothesis significance testing.[32] The core objective of the American Society of Statisticians is to progress research beyond 'all or nothing' hypothesis tests, which may be particularly important if the theoretical predictions within a study are weak.[30]

Clinical decisions *should not* be made solely on a p-value.[32] Many of the positive statistically significant conclusions from high-quality RCTs in sports physical therapy are probably no more than suggestive. Researchers must also understand that null hypothesis significance testing is only designed to work efficiently in the context of long-run repeated testing (exact replication).[30] A single significant result should not be

concluded as a “scientific fact.” The result should be interpreted as something worthy of further investigation,[12, 31] particularly if it was derived from a secondary outcome.

There is no consensus on how best to communicate results of testing scientific hypotheses. RCTs in orthopedics and sports medicine have traditionally used a frequentist approach based on deductive inference. Our calculation of FPR involved application of Bayes’ Theorem, where the central tenet is to consider how current data alter our “prior probability”, to generate a new, “posterior probability.” We initially used a “non-informative” prior probability of 50%, meaning that we assumed an even odds of treatment effect. As we audited clinical studies from a diverse field, there may be situations when hypotheses are more skeptical or optimistic. Therefore, we calculated FPRs based on both low [P(H1) =0.2] and high [P(H1) =0.2] prior probabilities. As expected, when prior probabilities were shifted closer to zero, the FPR was inflated; when we assumed a high prior probability of effect, 75% of findings had FPRs <8%.

There continues to be debate around the relative merits of a frequentist and Bayesian approach to statistical analysis. Our findings highlight how Bayesian thinking and conditional probabilities can affect the interpretation of null hypothesis significance testing.[4] For example, a statistically significant finding generated from a RCT examining the effects of jugular vein compression devices[29] on concussion incidence in contact sports (skeptical prior) should be interpreted with more caution than a statistically significant finding from a RCT testing the analgesic effects of topical cooling after a musculoskeletal injury (optimistic prior). In effect, Bayesian logic ensures that the

skeptical prior example requires more 'extreme' data before treatment effectiveness can be concluded. In contrast, the traditional frequentist approach, does not differentiate between these two research questions.

A key limitation of Bayes' Theorem is the uncertainty when determining what a suitable prior probability should be. One solution is a reverse Bayesian approach,[25] where the observed p-value is used to calculate the prior probability required to achieve a specific or minimal false positive risk (eg. 5%). This approach allows the researcher to determine whether the calculated prior probability is plausible or not. It has been suggested that 0.5 (or a 50:50 chance of success) might be the largest prior probability that can be legitimately assumed.[5] In our analysis, approximately 60% of positive (statistically significant, $p<0.05$) outcomes would require prior probabilities greater than 0.8 to achieve FPRs of 5%. Such extreme prior probabilities are likely unacceptable as they represent situations where a researcher is almost certain of treatment success (a non-zero effect), before the experiment is even initiated.

Trials with positive outcomes are published more often, and more quickly, than trials with negative findings.[16] The proportion of positive results in published scientific literature may be as high as 86%.[9] In our analysis of high-quality RCTs within sports physical therapy, we found an equal ratio of trials reporting positive and null effects. Although this might suggest that publication bias is not an issue within the sports physical therapy field, there were no trials reporting negative or harmful effects of an intervention. There may also be publication bias in lower quality studies, which we excluded. Trial registration is

considered an effective way to control publication bias,[20] and can help to prevent cherry-picking statistically significant results later. We found that only 29% of sports physical therapy trials were prospectively registered. It is important that this figure eventually increases to 100%. A broader and more complex challenge is that often, many trials have discord between the original registry data and the published data, despite registration.[11] Additional solutions have been proposed including: improved CONSORT compliance, from both researchers and editorial boards, and improvement to the post-publication peer review process. [11]

The evidence base for orthopaedics and sports medicine has been criticized for inappropriate participant selection[3] and high risk of bias.[22] Issues related to undefined primary endpoints and multiple comparisons have plagued the literature,[22] but their relevance has been difficult to quantify. Our results suggest that methodological shortcomings may be leading researchers in orthopaedics, sports medicine and sports physical therapy astray in their conclusions, and negatively influencing evidence-based practice.

Limitations

A recent audit of the PEDro database (The Physiotherapy Evidence Database (PEDro; http://www.pedro.org.au)) listed over 23 049 RCTs, of which 1098 have been undertaken in sports-related disciplines.[19] We limited inclusion to RCTs archived within the PEDro database and used a cut off of >6/10 (on the PEDro scale) to define high quality. Our audit was limited to results from single experiments and we did not fully consider false

discoveries relating to other important sources such as the use of multiple treatment arms, analysis of multiple outcomes, and multiple analyses of the same outcome at different times.[21] FPR is likely to increase if lower quality methodological designs are employed,[5] therefore our FPR estimations are likely conservative in the broader context of all clinical trials. We did not focus on false negative findings or outcomes deemed to be surrogate in nature (e.g. biomarkers). We acknowledge the importance of directing future work in this area; our primary focus was on the risk of false positive findings regarding outcomes that reflect real-clinical settings.

Recommendations for future research

Future reports should include exact figures for p-values rather than thresholds ($p<0.05$) and avoid using the term significant.[4] We were often unable to calculate FPR due to missing data. It is essential that researchers accompany reported p-values with effect sizes, corresponding confidence intervals, and ideally a minimum false positive risk estimation. It is important that there is a continued focus on the mandatory preregistration of study protocols, publication of pre-study power calculations and effect sizes, including any negative findings.

While the proper use of statistics will help to minimize false discoveries in research, there are other factors currently influencing the risk of erroneous findings in the sports physiotherapy field. It is possible that the existing academic system in sports physical therapy (like many other areas of healthcare) might increase the risk of erroneous or selective publishing, because career milestones such as promotion or tenure are often

297 determined by the volume of researchers' publication record.[13] Journal editors, reviewers
298 and grant-review committees may also favor scientific findings that are confirmatory, clear
299 and complete[2] — limiting the chances of disseminating negative or contradictory research
300 findings. We encourage researchers to examine FPR in other disciplines of health care.
301
302 To calculate FPR, we used an online calculator that uses post-hoc statistical power to
303 inform FPR values. It is possible that some studies recorded very large effect sizes due
304 to sampling variation, which consequently overestimates statistical power (a posteriori)
305 and potentially inflates the FPR estimate. Future research could include additional FPR
306 estimations using a range of statistical power parameters (partially post hoc power).[8]
307
308 **Conclusion**
309 Research conclusions should not be based solely on Null Hypothesis Significance Testing
310 (NHST) and p-values. Over 60% of statistically significant findings ($p<0.05$) reported in
311 the physiotherapy literature, carried FPRs greater than 5% and the median FPR was 9%
312 (assuming a prior probability of 0.5).
313
314 **Ethics approval and consent to participate**: Not applicable
315 **Consent for publication**: Not applicable
316 **Availability of data and material**: Data sharing is not applicable to this article as no
317 datasets were generated or analyzed during the current study
318 **Competing interests**: The authors declare that they have no competing interests"
319 **Funding**: Nil

Authors' contributions: All authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript. CB and JS were involved in the concept, design and writing. CB and JR were involved in the analysis. All authors were involved in final submission and revision of the manuscript.

Acknowledgements: Not applicable

Key points

Findings

Many of the positive statistically significant conclusions from high-quality RCTs in sports physiotherapy are probably no more than suggestive. We estimate the median false positive risk (FPR) in this field to be 9% (25^{th}-75^{th} percentile: 2-24%).

Implications

Research conclusions should not be based solely on Null Hypothesis Significance Testing (NHST) and p-values. The risk of making a false claim of treatment effectiveness can be reduced through, more rigorous consideration of pre study odds (ie. the chances that a treatment will work a priori) and reporting of FPR (a posteriori).

Cautions

This audit was limited to high quality, 2-arm RCTs. We also did not consider other sources of false discoveries in research such as: the use of multiple treatment arms, analysis of multiple outcomes, and multiple analyses of the same outcome at different time points.

References

1. Baker M. Reproducibility crisis: Blame it on the antibodies. *Nature.* 2015;521(7552):274-276.
2. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature.* 2012;483(7391):531-533.
3. Bleakley C, MacAuley D, McDonough S. Are sports medicine journals relevant and applicable to practitioners and athletes? *Br J Sports Med.* 2004;38(5):E23.
4. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci.* 2014;1(3):140216.
5. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. *R Soc Open Sci.* 2017;4(12):171085.
6. Colquhoun D. The False Positive Risk: A Proposal Concerning What to Do About *p*-Values. *The American Statistician.* 2019;73:192-201.
7. de Morton NA. The PEDro scale is a valid measure of the methodological quality of clinical trials: a demographic study. *Aust J Physiother.* 2009;55(2):129-133.
8. Dziak J, Dierker L, Abar B. The interpretation of statistical power after the data have been gathered. *Current Psychology.* 2019.
9. Fanelli. Negative results are disappearing from most disciplines and countries. *Scientometrics.* 2012;90(3):891-904.
10. Freedman LP, Cockburn IM, Simcoe TS. The Economics of Reproducibility in Preclinical Research. *PLoS Biol.* 2015;13(6):e1002165.
11. Goldacre B, Drysdale H, Dale A, et al. COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials.* 2019;20(1):118.
12. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol.* 2008;45(3):135-140.
13. Grimes DR, Bauch CT, Ioannidis JPA. Modelling science trustworthiness under publish or perish pressure. *R Soc Open Sci.* 2018;5(1):171511.
14. Harris JD, Cvetanovich G, Erickson BJ, et al. Current status of evidence-based sports medicine. *Arthroscopy.* 2014;30(3):362-371.
15. Heneghan C, Goldacre B, Mahtani KR. Why clinical trial outcomes fail to translate into benefits for patients. *Trials.* 2017;18(1):122.
16. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev.* 2009(1):MR000006.
17. Ioannidis JP. Why most published research findings are false. *PLoS Med.* 2005;2(8):e124.
18. Kamper SJ. Interpreting Outcomes 2-Statistical Significance and Clinical Meaningfulness: Linking Evidence to Practice. *J Orthop Sports Phys Ther.* 2019;49(7):559-560.
19. Kamper SJ, Moseley AM, Herbert RD, Maher CG, Elkins MR, Sherrington C. 15 years of tracking physiotherapy evidence on PEDro, where are we now? *Br J Sports Med.* 2015;49(14):907-909.
20. Laine C, Horton R, DeAngelis CD, et al. Clinical trial registration: looking back and moving ahead. *Lancet.* 2007;369(9577):1909-1911.
21. Li G, Taljaard M, Van den Heuvel ER, et al. An introduction to multiplicity issues in clinical trials: the what, why, when and how. *Int J Epidemiol.* 2017;46(2):746-755.
22. Lohmander LS, Roos EM. The evidence base for orthopaedics and sports medicine. *BMJ.* 2015;350:g7835.
23. Longstaff C, Colquhoun D. http://fpr-calc.ucl.ac.uk/. Accessed 01-02-2019.
24. Marino MJ. How often should we expect to be wrong? Statistical power, P values, and the expected prevalence of false discoveries. *Biochem Pharmacol.* 2018;151:226-233.

25. Matthews R. Why should clinicians care about Bayesian methods? *J Stat Plan Inference.* 2001;94:43-58.

26. Moseley AM, Elkins MR, Janer-Duncan L, Hush JM. The Quality of Reports of Randomized Controlled Trials Varies between Subdisciplines of Physiotherapy. *Physiother Can.* 2014;66(1):36-43.

27. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science.* 2015;349(6251):aac4716.

28. Ramagopalan S, Skingsley AP, Handunnetthi L, et al. Prevalence of primary outcome changes in clinical trials registered on ClinicalTrials.gov: a cross-sectional study. *F1000Res.* 2014;3:77.

29. Smoliga JM, Wang L. Woodpeckers don't play football: implications for novel brain protection devices using mild jugular compression. *Br J Sports Med.* 2018.

30. Szucs D, Ioannidis JPA. When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment. *Front Hum Neurosci.* 2017;11:390.

31. Wood J, Freemantle N, King M, Nazareth I. Trap of trends to statistical significance: likelihood of near significant P value becoming more significant with extra data. *BMJ.* 2014;348:g2215.

32. Yaddanapudi LN. The American Statistical Association statement on P-values explained. *J Anaesthesiol Clin Pharmacol.* 2016;32(4):421-423.